Machine Learning for Prediction and Analysis of Public Financial Data

Team Members: Rajat Ahuja, Casey Hu, Seungjun Lee, Aaron North (Mentor: Dr. Haris Vikalo)

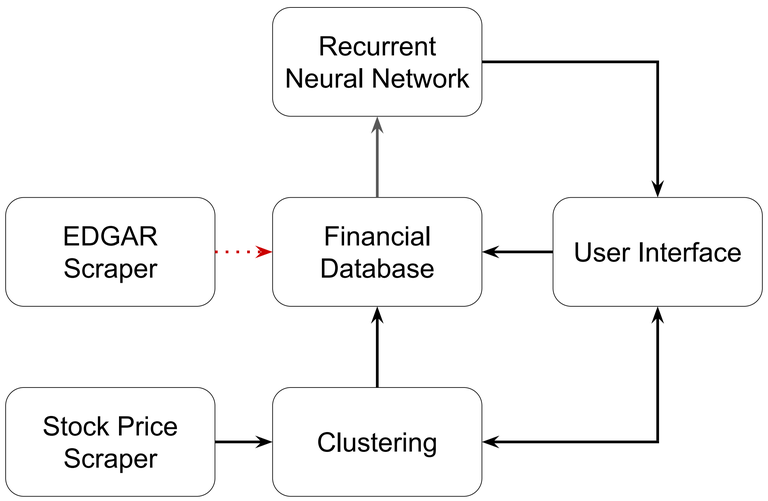
Problem Background

Predicting future trends in the stock market has been a long-researched topic in market analysis. State Street introduced to us the SEC database known as EDGAR (Electronic Data Gathering, Analysis, and Retrieval System), which contains huge amounts of financial data that has not been used to its fullest potential in the field of stock market prediction. **The aim of our project was to implement a scraper for EDGAR data and to create a web application that uses that data in a machine learning model to make correlations and predictions on future values of stocks.** This tool could be used to bolster State Street's advisory services.

System Requirements

- System will be a web application that receives financial data as input and outputs a prediction for companies' future stock value.
- Starter data will be scraped from the publicly available quarterly reports in EDGAR.
- Users should be able to upload any number of companies' data as a CSV file.
- The back end machine learning model will train on a Pandas DataFrame with each feature vector labeled with the stock price.

Design Summary



- System initially contains financial data from 9 different companies.
- Users can import financial data of additional companies in CSV format.
- The website clusters the companies using their stock price history from Alpha Vantage API.
- The website uses the clustering results to build a model for the queried company.
- The website outputs a prediction of the queried stock's value in the next quarter.
- The RNN is created using TensorFlow and utilizes LSTM cells, Adam optimizer, and the mean-absolute-error loss function.
- The network consists of one hidden layer with 12 nodes in the first layer.

Testing & Evaluation

- **Methods:** Predicted the stock prices for each company in the third quarter of 2019 using data from 2008 to the second quarter of 2019. Error was measured as the percentage deviation of the predicted value from the actual value.
- **Improvements:** Added more rows of data to enable the model to have more memory in the LSTM cells. Also reduced overfitting by only training using stock data from companies within the same cluster.
- **Results:** Predictions for all companies passed a sanity check (i.e., they were within the same order of magnitude). The average error when validating on the third quarter of 2019 utilizing data from our nine starter companies was **22.38%**
- **Conclusions:** As we had relatively few rows of data per company, minimizing the number of hidden layers and increasing the number of epochs produced the lowest validation error. Batch size had no significant effect on reducing the error. Our system can only predict one quarter into the future with minimal error. We could expand our model with more data and greater memory of the LSTM cells, but it would be difficult to accurately predict values further in the future due to the volatility of the stock market.